Genomics 92 (2008) 134–143

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Structural differences of orthologous genes: Insights from human–primate comparisons

Tuan Meng Lee [a,b,c], Leonard Lipovich [a,b,*]

[a] School of Computer Engineering, Nanyang Technological University, Singapore
[b] Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore
[c] Alexandra Hospital, National Healthcare Group, Singapore

ARTICLE INFO

Article history:
Received 26 September 2007
Accepted 2 May 2008
Available online 7 July 2008

Keywords:
Comparative genomics
Intergenic splicing
Bioinformatics
Primates
Gene structure
cDNA

ABSTRACT

The genomic basis of phenotypic distinctions between humans and nonhuman primates remains insufficiently explained. We hypothesized that interspecies structural differences of orthologous genes can cause such distinctions and searched protein-coding genes conserved between humans and nonhuman primates for species-specific initial and terminal exons. We inferred gene structure differences from genomic locations where portions of primate transcripts aligned with the human genome outside of any human exons. Of 22,466 high-confidence FANTOM3 human transcriptional units, 7424 (33%) had nonhuman primate full-length cDNA support. One hundred eighty-three of the loci contained 68,424 bp of sequence exonic in nonhuman primates but not humans. Fifty-four of 183 included species-specific portions of protein-coding regions. Six genes had evidence of intergenic splicing in a nonhuman primate but not in human. It is imperative that primate transcriptome projects be accelerated on par with genome projects to understand better interspecies gene structure distinctions.

The human interest in improving our health is intimately linked to the argument that certain organismal characteristics, including those relevant to disease, are uniquely human. That argument is bolstered by the numerous differences in pathology, anatomy, and behavior between humans and nonhuman primates. We are more susceptible than the great apes to specific diseases: AIDS, myocardial infarction, and hepatitis B/C complications [1]. Chimpanzee–human sequence identity surpasses 98% in alignable genomic regions [2]. Humans and chimpanzees are profoundly phenotypically different in many additional ways, but specific genetic features responsible for most of the differences are unelucidated.

Several approaches have been historically employed to explain interspecies phenotypic differences [1]. One has centered on comparative genomics of conserved regions. The mouse, a key model organism, is believed to have shared a common ancestor with humans 75–80 million years ago [3]. Due to neutral divergence and selective pressures, genomic differences have arisen since the primate–rodent divergence. For instance, low-copy repeats and segmental duplications have undergone lineage-specific accretion during primate evolution [4]. Accordingly, human–mouse comparisons can miss functional regions arising uniquely along the primate lineage [3].

While the evolutionary distance to mouse introduces limitations, comparisons of the human genome to closely related species can yield false positive results: nonfunctional regions appear conserved, as there is not enough time for mutation to occur. Phylogenetic shadowing can address this problem. Regions consistently conserved among multiple species are identified by this method and have a higher probability of being functional (see Fig. 1, [5]).

In-depth analysis of chromosomal changes, indels, repeats, duplications, and gene conversions has been used to study interspecies distinctions in genomic alignments. These distinctions at multiple loci contribute to differences between humans and nonhuman primates by affecting genes and regulatory regions directly [2,6–8].

Gene duplication, either segmental or retrotransposition-mediated, can create new genes with new biological functions. Duplicated genes can become nonfunctional (pseudogenes), neofunctional (acquire a new function), or subfunctional (adopt a portion of the previous function) [9]. Up to 30% of human segmental duplications have taken place since the human–chimpanzee common ancestor split [10]. There are 200–300 species-specific retroposed gene copies in humans and chimpanzees [2]. Both duplication types may be neofunctional in a species-specific manner.

Interspersed repetitive elements, especially the primate-specific *Alu* subclass of SINEs, have contributed profoundly to primate evolution. *Alu* insertions have affected the open reading frames of protein-coding genes [11], served as markers distinguishing primate species [12], and may even have facilitated the evolution of large brains in humans by inactivating the gene for a cell-surface glycoprotein that interacts with etiologic agents of bacterial meningitis [13].

Another existing perspective on phenotypic uniqueness entails searches for gene signatures of human-specific rapid evolution. High

* Corresponding author. Present address: Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201-1928, USA.
E-mail address: llipovich@med.wayne.edu (L. Lipovich).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.05.006

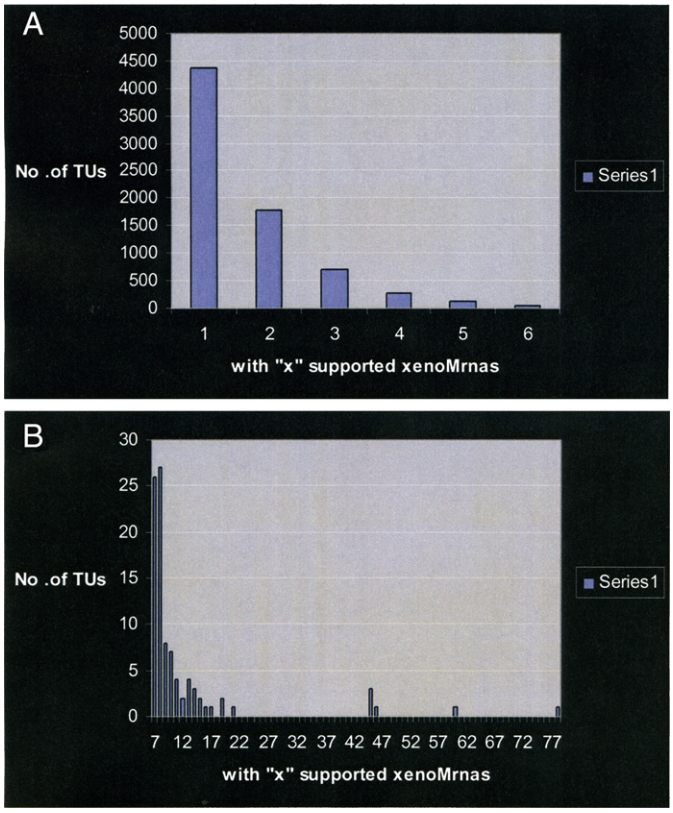

Fig. 1. Distribution of the number of nonhuman primate xenoMrnas per human TU. (A) TUs with six or fewer supporting xenoMrnas. (B) TUs with seven or more supporting xenoMrnas.

nonsynonymous-to-synonymous substitution (K_a/K_s) ratios are a property of genes undergoing accelerated evolution and positive selection [14]. An additional perspective focuses on non-protein-coding sequence changes. In a key early study, King and Wilson revealed that human and chimpanzee protein-coding sequences are highly similar. They proposed that conserved genes may be regulated differently between the two species, influencing phenotypic differences [15]. More recently, Glazko et al. found that 80% of proteins are different between the human and the chimpanzee, although the specific differences may be too minor to explain phenotypic distinctions. The phenotypic differences may be controlled by a few regulatory or major-effect genes [16].

Recent cDNA cloning projects and genome-tiling array experiments reveal that half or more of the transcriptome does not encode proteins [17,18] and suggest that noncoding RNAs (ncRNAs), including ones not conserved between species, merit consideration as contributors to species distinctions. Previously described ncRNAs (such as tRNAs, rRNAs, spliceosomal RNAs, catalytic RNAs, and small RNAs) play essential roles, suggesting that other ncRNAs can also be functional. From 512 FANTOM2 mouse ncRNA sequences conserved in human, 8 functional ncRNAs were identified [19]. However, most ncRNAs are not conserved, which does not imply lack of function [20], because metrics other than exon conservation may be more appropriate gauges of these RNAs' functionality [21]. Medically important human ncRNAs include DLEU2, 7H4, and BIC, implicated in lymphocytic leukemia, postnatal development, and Hodgkin lymphoma, respectively [3].

NcRNAs exert regulatory effects by hybridizing to mRNA molecules, as *trans*- and *cis*-antisense transcripts (in prokaryotes and eukaryotes) and as microRNAs in eukaryotes, which favor 3′ untranslated regions (3′UTRs). Therefore, a better understanding of gene structure differences, including at UTRs, may help understand species-specific gene expression regulation by ncRNAs.

Interspecies comparisons of gene expression levels represent another approach used to link genomic and phenotypic distinctions, as it is difficult to predict changes in gene expression purely by comparing genomic sequences [22]. Microarrays based on human sequences may not always detect expression changes or alternative splicing in nonhuman primates, due to human–primate divergence at probe-homologous regions [23]. Moreover, expression differences can be due to neutral evolution [24]. Finally, many functional products are not proteins, but result from their enzymatic activity (e.g., lipids, glycans, and bioactive small molecules). Thus, gene expression studies should be complemented by other approaches, including lipomics and glycomics [1].

Entirely primate-specific genes, while rare, constitute yet another potential explanation for human–primate phenotypic differences. De novo birth of brain-expressed genes has taken place in the hominid lineage by complex mutational, antisense, and retropositional mechanisms [25], and primate-specific repetitive sequences have contributed to exons of new genes with brain expression [26], leading to conclusions that gene genesis has played a role in the accretion of primate evolutionary novelties [27].

Several assumptions have historically dominated genetics [28]. These include the one gene–one protein rule, the Central Dogma, and the deterministic perspective on gene expression. Venter et al. reintroduced the term "transcriptional unit (TU)," originally used in the 1970s to describe ribosomal-RNA operons, to refer to gene models, including those distinct from known genes, supported by multiple levels of evidence, including cDNAs and ESTs [29,30]. We define a TU as a collection of all known transcripts sharing any exonic sequences on the same strand in the same locus [31].

We have coupled a published human TU catalog to a novel framework for enumerating human–primate distinctions. The limitations of existing perspectives on interspecies differences inspired us to examine a class of genomic distinctions that has rarely, and usually in single-gene case studies, been investigated to date: orthologous gene structure differences evident from aligning human and nonhuman primate cDNAs to human genomic regions. This approach enabled us to utilize all publicly available primate transcriptome sequences, including those from species lacking sequenced genomes. Gene structure differences thereby revealed can be phenotypically relevant because they indicate unique protein-coding DNA sequence (CDS) and UTR subsequences of orthologous human and primate genes. The unique subsequences may result in species-specific protein portions and species-specific regulation by mRNA-binding ncRNAs and/or RNA-binding proteins, respectively.

The goal of the present study was to search for initial-exon and terminal-exon gene structure differences at orthologous loci between humans and nonhuman primates, using transcript-to-genome alignments. Attaining this goal, we compiled a quantitative catalog of orthologous gene structure differences that involve initial or terminal exons or exon fragments unique to nonhuman primates.

Results

Only a subset of human TUs is supported by nonhuman primate cDNAs

The initial high-confidence FANTOM3 [33] TU dataset comprised 22,466 TUs, approximately 59% of the 38,037 total FANTOM3 TUs that had been successfully mapped to the HG18 human genome assembly. The remaining 41% were not considered because they were not supported by human cDNAs or multiple human ESTs. We matched the 19,699 nonhuman primate GenBank cDNAs (Supplementary File "distinct nonhuman primate cDNAs.xls") mappable to the HG18 assembly with specific human TUs mappable to that assembly, capturing the species origin of each nonhuman primate cDNA in the process. The genomic coordinate ranges of 6893 nonhuman primate cDNAs, when mapped to the human genome, did not overlap any

coordinates of high-confidence human TUs. These nonhuman primate artifacts or genes absent from the human TU dataset were not pursued further. Of the initial dataset, 7424 (33%) TUs were supported by 12,806 nonredundant nonhuman primate cDNA accessions (Fig. 1; Supplementary File "high confidence TUs w/ supporting xenomrna if any.xls"). The results indicate that the majority of human TUs currently lack nonhuman primate cDNA support and highlight a need for deeper sequencing of primate transcriptomes. Such sequencing would enable more thorough human–primate ortholog gene comparisons that exploit actual exon–intron structures of expressed genes, not just genomic sequence alignments.

Nonhuman primate cDNAs reveal candidate terminal-exon differences at orthologous loci

For 2488 nonhuman primate cDNAs (listed in Supplementary File "distinct nonhuman primate cDNAs.xls" as [largetstart,largetend]), both the start and the end of each cDNA were >1 kb distant from the start and end of any human TU reference transcript. These primate cDNAs might reflect mapping artifacts, correspond to genes not supported by high-confidence human TUs, and/or have abnormally long human genomic spans because of cDNA library chimerism, and hence they were not analyzed further.

A set of 5371 (listed in Supplementary File "distinct nonhuman primate cDNAs.xls" as [smalltstart,smalltend]) nonhuman primate cDNAs had both start and end aligning within 1 kb of the start and end, respectively, of a high-confidence human TU. These primate cDNAs most likely reflected expression from primate orthologs of the human TUs. But since they lacked evidence of major structural differences in the 5′- and 3′-end positions between primates and humans, they were also not considered further.

Elimination of the 2488 primate cDNAs with large differences at both ends relative to human genes and the 5371 primate cDNAs that recapitulate human gene structures without terminal differences resulted in 5250 prospective cases of orthologous-loci terminal-exon differences. Of the 5250 cases, 4811 involved a primate cDNA with <1 kb difference in mapping to the human genome at one end relative to the orthologous human cDNA and a difference of between 1 and 100 kb at the other end. We did not analyze those cases, even though they might represent a pool of bona fide interspecies differences useful in future studies; we anticipated that differences of greater than 100 kb presented greater potential for confirmation as readily evident, and major, orthologous-gene structure distinctions. Therefore, we manually annotated only the cases (Supplementary File "list of 439 orthologous pairs.xls") in which one end of a nonhuman primate cDNA, mapped to the human genome, was <1 kb away from the end of the orthologous human gene, while the other end of that nonhuman primate cDNA was >100 kb away from the other end of that human gene.

Of the 439 cases, 208 were eliminated: 171 lacked terminal-exon interspecies differences upon manual inspection in the UCSC Genome Browser [33] or belonged to known TCR and MHC loci whose repetitive nature made interpretation difficult, and 37 had ambiguous multiple genomic mappings. A further 2 duplicate records and 9 instances of putative intergenic splicing, to be discussed further, were also separated (Supplementary File "list_of_439_orthologous_pairs.xls"). Hence, we extracted 220 individual nonhuman primate cDNAs, which corresponded to 183 unique human genes (Supplementary File "list of distinct human accession nos that feature structural differences at terminal ends.xls") for subsequent analysis.

Sequences transcribed uniquely in nonhuman primates contribute to protein-coding and untranslated regions of conserved genes

To weigh the relative contributions of species-specific exons of orthologous genes to protein sequence, vs via UTRs, we quantitated two metrics: number of nonhuman primate cDNAs that had species-specific UTR and/or CDS and amounts (bp) of CDS and UTR sequence contained in those primate-specific terminal exons that were exclusive to nonhuman primate cDNAs (Tables 1A, B).

We first considered 220 nonhuman primate cDNAs corresponding to all 183 genes with evidence of species-specific terminal exons (Supplementary File "categorisation of unique terminal exons.xls"). One hundred fifty of the 220 nonhuman primate cDNAs had species-specific regions that contained only UTR sequence. The remaining 70 contained at least some protein-coding sequence. Therefore, most frequently, exonic sequence unique to a nonhuman primate cDNA was confined to a UTR (Table 1A). In order of decreasing total affected sequence length, 5′UTR bases were most frequently affected by cDNA-supported terminal-exon differences between humans and nonhuman primates. They were followed by 3′UTRs and then by CDS (Table 1B).

A subset of species-specific terminal exons in primates is completely devoid of human cDNA and human EST support

The 183 human genes with species-specific terminal exons in orthologous primate cDNAs had primate-specific terminal exons absent from the majority of, but not necessarily all, orthologous human cDNAs and ESTs. Therefore, some of those primate terminal exons provided confirmation that specific minor-frequency human alternative splicing events affecting terminal exons occurred at orthologous primate genes, but did not provide evidence for primate-specific exons completely unused in humans. To distinguish between those two possibilities, we defined a subset of human/primate terminal-exon differences completely unsupported by human cDNA/EST data.

Fifty-five human genes (Supplementary File "categorisation of 55 superhigh confidence TUs.xls") had nonhuman primate orthologous cDNAs with terminal-exonic sequences completely devoid of human cDNA and EST support. We refer to these 55 genes as the "super-high-confidence" subset of the 183. Because of the much greater depth of human vs nonhuman transcriptome coverage in public cDNA/EST databases, it is unlikely that the terminal exons observed in nonhuman species for those 55 genes are used in any orthologous human transcripts.

The relative proportions of species-specific sequences corresponding to translated and untranslated regions, as observed in all 183 genes, were recapitulated in these 55 genes. Specifically, the 5′UTR category still ranked first and the CDS category still ranked last; the number of

Table 1A
Categorization of unique structures found at terminal exons of nonhuman primates based on individual nonhuman primate transcripts

UTR and CDS content of primate-specific exons absent in human	Number of distinct nonhuman primate cDNAs
5′UTR only	99
CDS only	10
3′UTR only	51
5′UTR and CDS	41
CDS and 3′UTR	10
5′UTR, CDS, and 3′UTR	9
Total	220
55 super-high-confidence TUs	
5′UTR only	31
CDS only	0
3′UTR only	13
5′UTR and CDS	5
CDS and 3′UTR	0
5′UTR, CDS, and 3′UTR	6
Total	55

Table 1B
Number of base pairs of unique primate-specific exonic sequences

	5′UTR	CDS	3′UTR	Total No. of bp
TUs	28,586	14,152	25,686	68,424
55 super-high-confidence TUs	7,264	3,307	5,358	15,929

base pairs of species-specific CDS in nonhuman primates was still roughly half the number of base pairs of species-specific 5′UTRs (Table 1B). However, there was a reduction in the absolute number of nonhuman primate cDNAs whose species-specific terminal exons affected the CDS (Table 1A). Therefore, UTRs detected by our analysis as unique to nonhuman primate cDNAs may be more likely to be completely devoid of human cDNA and EST support, while CDS regions detected by our assessment are more likely to be present in minor-frequency splice variants of the corresponding human orthologs.

Nearly half (26) of the nonhuman primate genes whose unique species-specific exons lacked any human cDNA or EST support are involved directly in biological or evolutionary aspects of primate uniqueness (defined to include cancer, brain function, neurodegenerative disease, immunity, glycoproteins, reproduction, rapid evolution, and positive selection) (Supplementary File "gene names of 55 superhigh confidence TUs.xls"). Specific genes included two relevant to HIV and SIV infection (APOBEC-binding HNRPAB and NFAT5), a brain size gene mutated in neurodegenerative disorders and known to be affected by segmental duplications and structural polymorphism (myomegalin), and multiple glycoprotein biosynthesis enzymes, important because cell-surface glycoprotein differences are prominent in primates [34]. Thirteen of the 55 genes, including 2 expressed in brain, were involved directly in cancer pathways, suggesting that primate-specific gene structure differences at cancer loci, in addition to coding-sequence differences between orthologous human and chimpanzee cancer genes [35], may be relevant to the differential incidence of cancers between primate species.

Specific protein functions are putatively enriched in human genes with terminal-exon differences relative to primates

To test for enrichment or depletion of specific biological functions within a subset of genes, gene ontology classifications can be retrieved for all genes in the set, and the frequencies of encountering each classification can be compared with those in the entire human gene catalog. We hypothesized that genes encoding specific biological functions potentially relevant to human–primate phenotypic distinctions may be enriched among our 183 genes with human–primate terminal-exon differences and/or in our super-high-confidence subset of 55 of those genes. To test this hypothesis, we subjected both gene sets to enrichment and depletion analysis using two robust controlled-vocabulary ontology systems, PantherDB and GO, with multiple-testing correction. Three independent tools were used for GO enrichment analysis [36–39].

Specific processes involving protein interactions, including protein depolymerization, posttranslational modifications, cytoskeletal functions, and calcium binding, were enriched in the 183-gene set and consistently detected by all four tools at $P<0.015$ (Table 2). This contrasts with historical expectations of primate-specific functions limited to neuronal activities, synaptogenesis, reproduction, and immunity. Specific protein-interaction-related processes might be relevant to phenotypic differences between humans and primates because of their link to HIV infection: interactions of host and HIV proteins influence the dimerization of the transcriptase (p51/p66), while cell-fusion processes are important for viral spread and therefore are potential targets in drug design [40]. We speculate that the distinctions between the human immune response to HIV and the chimpanzee immune response to SIV might be encoded by structure differences at orthologous genes in the functional categories detected as enriched in our unique-terminal-exon orthologous-pair dataset.

Overrepresentation of PantherDB categories describing neuronal and synaptic functions was observed in our starting datasets (the 22,466 high-confidence human TUs and the 7424 human TUs supported by nonhuman primate cDNAs). This was likely due to a bias in favor of genes with these functions in human TU and primate cDNA datasets, as brain and neuronal transcriptome libraries might be sequenced more deeply or more frequently than other cDNA libraries.

Novel intergenic splicing events unique to primates are suggested by nonhuman primate cDNA alignments to the human genome

We detected six cases of intergenic splicing supported by nonhuman primate cDNA data (Fig. 2A). Intergenic splicing refers to the joining of exons from genomically adjacent but biologically separate genes within a single mRNA and is rarely observed in mammals [41]. A representative instance of intergenic splicing unique to the nonhuman primate cDNA set is illustrated in Fig. 2B. All cases except one (CYP2C19–CYP2C9 [42]) involve genes not previously known to be intergenically spliced. Because intergenic splicing as defined here is internal rather than terminal, relative to a gene structure, and involves multiple exons from both genes, these six loci are not included in our main dataset of human–primate terminal-exon distinctions. Absence of human cDNA or EST support for the five novel intergenic splicing events from nonhuman primates, despite the much greater depth of human vs nonhuman transcriptome coverages in public databases, suggests that these events are completely absent in humans, although the possibility that they are extremely rare in humans cannot be formally excluded. The GT–AG canonical introns spliced intergenically between the adjacent genes in each case argue strongly against a cloning artifact or chimerism explanation for these transcripts.

Discussion

Functional implications of terminal-exon differences at orthologous protein-coding loci between humans and primates

Mapping cDNA/EST-derived transcriptome sequences to genome assemblies has generated a wealth of gene structure data. However, structural differences among primates at known conserved protein-

Table 2
Classification of human genes that have terminal-exon structure differences demonstrable by nonhuman-primate cDNA comparisons, using a variety of tools, shows that protein-related activities are consistently overrepresented based on a cutoff, adjusted P value <0.015

Tool	Process	Adjusted P
David (http://david.abcc.ncifcrf.gov/summary.jsp)	Protein binding	5.6×10^{-6}
FuncAssociate (beta version) (http://llama.med.harvard.edu/cgi/func/funcassociate)	Negative regulation of protein metabolic process	<0.001
	Protein depolymerization	<0.001
Genecodis (http://genecodis.dacya.ucm.es/)	Protein binding	1.28×10^{-2}
Panther (http://www.pantherdb.org)		
Reference list: human AB1700 genes	Protein phosphorylation	4.69×10^{-3}
	G-protein modulator	1.04×10^{-3}
	Select regulatory molecule	4.77×10^{-3}
	Membrane traffic protein	1.48×10^{-2}
	Select calcium-binding protein	1.25×10^{-2}
Reference list: NCBI H. sapiens genes	Protein phosphorylation	9.10×10^{-3}
	G-protein modulator	3.64×10^{-3}
	Select regulatory molecule	1.31×10^{-2}

Accessed on August 23, 2007. "Unclassified" functions are omitted.

A

Accession No. (Nonhuman-primate)	Organism	Tissue	Accession No. (Human)	Intergenically spliced products	Intergenic introns
CR857653	Pongo pygmaeus	Kidney	AF390175	MIA2-CTAGE5	GT…AG (5th intron from left)
CR861367	Pongo pygmaeus	Cortex	BC018125	UCHL3-LMO7	GT…AG (5th intron from left)
AM048758	Pongo pygmaeus	Cortex	BX537684	MSTO1-LST005	GT…AG (4th from right)
AM048758	Pongo pygmaeus	Cortex	BC070067		GT…AG (4th from right)
CR861351	Pongo pygmaeus	Cortex	BX537684		GT…AG (4th from right)
CR861351	Pongo pygmaeus		BC070067		GT…AG (4th from right)
CR861418	Pongo pygmaeus	Cortex	BC024315	LCOR-C10ORF12	GT…AG (2nd from right)
DQ074807	Macaca fascicularis	N/A	M61855	CYP2C18-CYP2C9 [31]	GT…AG (1st from right)
AB173643	Macaca fascicularis	N/A	CR596160	ZFP106-TMEM87A	GT…AG (1st from right)

B

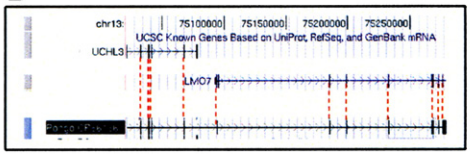

Fig. 2. Intergenic splicing in nonhuman primates. (A) Cases of intergenic splicing supported by nonhuman primate cDNA data. (B) A representative instance of intergenic splicing in nonhuman primate cDNA data without human cDNA or EST support. *Note.* The orangutan cDNA CR861367 contains exons from two genes (tracked by vertical dotted lines). This fits the definition of intergenic splicing (joining of exons from adjacent but separate genes).

coding genes have not previously been inspected on a genomewide scale.

We addressed this problem by comparing human transcriptome/human genome alignments with orthologous nonhuman primate transcriptome/human genome alignments. We searched for gene structure differences between humans and nonhuman primates, because these differences may contribute to species-specific phenotypes. In contrast, existing studies have emphasized genomic sequence, not transcript, comparisons. An advantage of our approach is that we are able to infer interspecies gene structure differences from nonhuman primate cDNAs, even if the genomes of those species have not been sequenced.

A diverse range of structural differences exists at orthologous loci between humans and nonhuman primates, including terminal-exon differences detected by our algorithm. These distinctions—in addition to interspersed repeats, positive selection, and segmental duplications—may account for some phenotypic differences between humans and other primates. Our approach establishes a foundation for thorough, quantitative assessments of conserved genomic sequences at orthologous loci that are incorporated into species-specific exons. By enumerating known genes affected by structural differences at their termini, we enable construction and testing of hypotheses regarding the impact of specific ortholog structure differences on phenotypic uniqueness.

Genomewide trends of 5′, 3′, UTR, and CDS participation in species-specific primate gene structures probably cannot be inferred from our small dataset. We were also unable to make functional conclusions unique to specific narrow phylogenetic groupings of primates, simply because the number of human mRNAs in GenBank exceeds by sixfold the number of all nonhuman primate mRNAs combined. However, despite the limited transcriptome data, we were still able to show structural differences at genes that previously were not known to be structurally variant within primates, and to put a lower bound on the number of UTR and CDS bases affected by human/primate differences.

Six nonhuman primate transcripts are inferred to originate from intergenic splicing—a phenomenon rarely observed in mammals and whose mechanism and function are not well understood. The corresponding human loci, despite numerous cDNAs/ESTs, have zero human cDNA/EST support for intergenic splicing. Hence, species-specific intergenic splicing may exist in nonhuman primates at these loci. When attempting to translate the ORFs of the intergenically spliced nonhuman primate cDNAs, we did not detect a potential for chimeric fusion-protein translation in the nonhuman primates. Intergenic splicing events may be a benign consequence of inefficient splicing or may exert a regulatory impact by omitting specific splice sites and truncating the ORF.

Our application of controlled vocabularies to elucidate functional biases within the subset of human genes having primate-specific terminal-structure differences was not particularly informative. Biases toward immune, behavior, and reproductive genes, while expected, were not shown. This may be due to the small size of the curated dataset or the preponderance of other types of human–primate differences (e.g., promoter substitutions, coding-sequence substitutions, or segmental duplications, but not terminal-exon structural differences) among genes implicated in previously characterized functional distinctions within primates. However, four enrichment analysis tools utilizing two controlled vocabularies suggested that specific

protein activity-related genes may be overrepresented in our dataset of terminal-exon differences. Additionally, a literature-based perusal of the 55-gene super-high-confidence catalog of species-specific terminal-exon changes in nonhuman primates pinpointed several immunity, brain-size, and cancer genes, as well as genes previously known to be subject to positive selection and structural variation. We recognize that a random sampling of 55 genes might produce a few members of these categories but it is nevertheless remarkable that the particular genes in our dataset are candidates for structural modification in primate evolution. Gene ontology and controlled vocabularies should be combined with individual-gene functional analyses for the most comprehensive possible assessment of primate-specific function enrichment in our subset of genes.

Ortholog structure comparisons of human genes and their counterparts in a single nonhuman primate, Macaca fascicularis

The macaque transcriptome has been sequenced more deeply than that of any other nonhuman primate, thereby affording an opportunity to perform gene structure comparisons systematically between human and a single nonhuman primate species. We have hence determined the distribution of all UCSC-mapped *M. fascicularis* cDNAs in our analysis (Supplementary File "Macaca_fascicularis_mRNA_counts.ppt"). A set of 9819 *M. fascicularis* full-length cDNAs from GenBank had unambiguous precomputed UCSC BLAT mappings to the HG18 human genome assembly. Of these, 3002 did not overlap any of the 22,466 high-confidence FANTOM3 human TUs as defined in our paper, while the other 6817 did overlap human TUs. Of the 6817 *M. fascicularis* cDNAs with human TU support, 290 had a <1-kb difference at the 5′ end but >100-kb difference at the 3′ end between human and nonhuman genomic localizations of gene boundaries, or vice versa, and also had greater human genomic spans for the macaque-to-human cDNA-to-genome alignment than for the human-to-human alignment, reducing the likelihood that the macaque transcripts would prove to be artifactually terminally truncated rather than reflect bona fide initiation or termination sites. One hundred forty-three and 42 macaque–human cDNA matches were included in our total and super-high-confidence sets, respectively, of orthologous loci with terminal-exon structure differences.

A high percentage of *M. fascicularis* testicular cDNAs were previously reported to display macaque-specific 5′UTR exons unsupported by human cDNA data [51]. In that particular study, no other species were analyzed; only 622 human–macaque cDNA pairs were detected (1 order of magnitude less than the 6817 macaque cDNAs matching human TUs in our study); no clarifications were made regarding the exact methodology for declaring an exon to be provisionally macaque-specific, and neither 3′UTRs nor intergenic splicing was considered. Nevertheless, in conjunction with our report, studies such as those by Osada et al. [51] provide a valuable integrated framework for cataloging the complete scope of species-specific terminal-exon gene structure differences in primates.

Future directions in human–primate ortholog gene structure comparisons

We focused on a limited set of genes with single-end structural differences between humans and primates and considered only the subset of orthologs with >100 kb separation of human and primate mappings for the discrepant ends. Extending annotation to the much larger set of genes with <100 kb separation between discrepant-end mappings is likely to uncover numerous additional examples of primate-specific incorporation of conserved genomic sequences into UTRs and CDSs. It may also help detect functional category enrichment or depletion signals in this class of genes.

One logical implication of our study is that gene structure differences between orthologs may exist in places other than terminal exons. During manual curation of terminal-exon differences, we occasionally observed internal exons arising from apparent species-specific exonification of conserved intragenic sequence at primate orthologs (an example is the AB171923–BC030199 macaque– human orthologous pair), but did not attempt to catalog this class of events systematically. Internal-exon structural differences might hold significant potential for phenotypic impact, as they are expected to affect primarily CDS rather than UTRs.

Our study does not utilize ab initio gene predictions or purely computational evidence of homology. All human–primate gene structure differences that we cataloged are supported by experimental evidence (full-length and EST sequencing of cDNA libraries). Nevertheless, additional experimental work may aid in validation and multispecies comparisons of the gene structure differences. RT-PCR and RACE validation of the 55 super-high-confidence ortholog distinctions may be an appropriate way to test whether the corresponding terminal exons are indeed never used in human transcripts, even in the same tissues and at the same developmental time points at which they are used in nonhuman primates. Similarly, for the six intergenic-splicing events, gene-specific nested RT-PCR in appropriate human tissues may elucidate whether equivalent splicing events are ever invoked in human.

The fundamental role of RNA-binding proteins in posttranscriptional gene regulation through 3′UTR binding has been well established [43]. Signal transduction pathways modulate gene expression not only through DNA-binding transcription factor activities but also through posttranscriptional mechanisms effected by RNA-binding proteins [44]. In addition, the paramount importance of endogenous microRNAs (miRNAs) as gene expression regulators has been expounded in a number of recent studies [45] indicating that miRNAs bind, with imperfect but detectable homology, to specific target sites in 3′ UTRs of protein-coding mRNAs. Therefore, species-specific UTR subsequences may result in species-specific miRNA-mediated regulation of conserved genes. Intraspecies polymorphisms in 3′UTRs that create phenotypically important mRNA-recognition sites are already known [46]. However, species-specific differences in orthologous UTRs that enable or preclude specific RNA-binding protein interactions with mRNAs have not been elucidated.

We have cataloged 54,272 bp of 5′UTR and 3′UTR sequence that is included in nonhuman primate transcripts but rarely or never used in their orthologous human counterparts. Of that sequence 12,622 bp belongs to primate UTRs with zero human cDNA and EST support, representing a conservative lower bound on the amount of species-specific UTR recruitment in nonhuman primates. These putative primate-specific UTRs can be mined by motif-finding tools for known RNA-binding protein recognition and microRNA hybridization sites. RNA-binding protein cognate motifs [47] and microRNA-binding motifs [48] have been characterized computationally, although complexity exists because of the impact of RNA structure on RNA-binding proteins [47] and nonoverlapping expression profiles of some microRNAs relative to predicted targets [48]. Our results can provide the foundation for a pilot project aiming to establish how conserved RNA-binding proteins, and/or microRNAs, may regulate orthologous genes in humans and primates in species-specific ways because of species-specific cognate sites in structurally variant UTRs.

A premise of our analysis is that, to be detected, a primate-specific terminal exon or exon extension must be mappable to the human genome assembly. Primate-specific terminal cDNA fragments not mappable to the human assembly would not have a human–primate mapping distance discrepancy metric attached to them. However, our analysis of human genes supported by orthologous primate cDNAs can be extended to capture primate-specific cDNA termini that do not map to the human assembly. Such sequences may contribute additional UTR and CDS fragments to the genomewide set of interspecies gene-structure differences.

The involvement of some functions detected by GO and PantherDB as enriched in our set of structurally different genes, such as cytoskeletal organization and calcium binding, in HIV pathogenesis may make it worthwhile to understand more about these processes in nonhuman primates' natural defense against retroviruses. The insights gained may aid in human drug development.

In the course of compiling human and primate cDNA datasets and cDNA–genome alignments, we developed several resources that should be broadly useful to the primate comparative genomics community. In particular, we validated mappings of all publicly available nonhuman primate cDNAs to the human genome (Supplementary File "distinct nonhuman primate cDNAs.xls") and constructed tables that link nonhuman and human GenBank accession numbers, gene names (Supplementary File "gene_names_of_55_superhigh_confidence_TUs.xls"), genome coordinates (Supplementary File "categorisation_of_unique_terminal_exons(b).xls"), and species descriptors (Supplementary File "distinct_nonhuman_primate_cDNAs.xls"). These tables can facilitate any analysis of human genes that requires retrieval of experimentally determined homologous transcript sequences from nonhuman primates.

In an era of growing methodological sophistication in primate comparative genomics, we were able to pinpoint intriguing gene structure distinctions between humans and nonhuman primates, including at several loci of known importance in primate evolution, by applying a pipeline that focuses on terminal-exon differences. We conclude that known protein-coding genes may harbor an underappreciated contribution to human uniqueness through terminal-exon distinctions that set humans apart from nonhuman primates.

Materials and methods

Identification of nonhuman primate full-length cDNAs alignable with the human genome

We used the following datasets (UCSC; http://genome.ucsc.edu) to associate a nonhuman organism with a specific cDNA alignable with the human genome:

- xenoMrna.txt.gz (HG18 version)
- gbCdnaInfo.txt.gz
- organism.txt.gz

In brief, we related the GenBank accession numbers of xenoMrna sequences (nonhuman mRNA sequences aligned with the human genome) to the Linnaean binomen of their source organisms, through the organism name field of the CdnaInfo file (Fig. 3). Then we extracted all records whose organism name/ID referred to a nonhuman primate. We obtained the list of known primates from the NCBI Taxonomy Browser.

Identification and mapping of high-confidence human TUs

We define high-confidence human TUs as those supported by at least one unambiguously mapped human cDNA and more than one unambiguously mapped human EST in the FANTOM3 dataset. We used the following FANTOM3 files:

- composite_mapping.txt.gz [32] (http://www.genereg.net/complex_loci/dataset/hg17_v05/)
- TU.txt.gz [32] (http://www.genereg.net/complex_loci/dataset/hg17_v05/)
- all_mrna.txt (http://genome.ucsc.edu)

We used the first two files to identify high-confidence human TUs, but their initial mappings were to the older HG17 human assembly. However, we decided to use the most recent (HG18) mappings of nonhuman primate mRNAs on the human genome. Therefore, we utilized all_mrna.txt to map FANTOM3's reference transcripts of the human TUs to HG18 through their GenBank accession numbers.

Identification of terminal-exon structure differences at orthologous human–primate gene pairs from full-length cDNA-to-genome alignments

We undertook a structure-centric, rather than a sequence-centric, approach to conservation. Specifically, we were interested in identifying whether orthologous genes have species-specific unique exons. Genomic conservation is not sufficient to identify such exons; evidence of transcription, from mRNA datasets, is required to identify a genomically conserved region as transcribed.

Many types of structural differences are possible between orthologous genes. These include terminal and internal exons specific to one species, intron size differences, repeat insertions, short indels, and differential splicing. Terminal-exon distinctions are easy to identify because they involve alignment edges rather than internal blocks.

As artifactual truncation of cDNAs at 5′ and 3′ ends during cDNA library construction may occur, we focused only on genes that had structural differences confined to one end of the alignment (<1 kb difference at the 5′ end but >100 kb difference at the 3′ end between

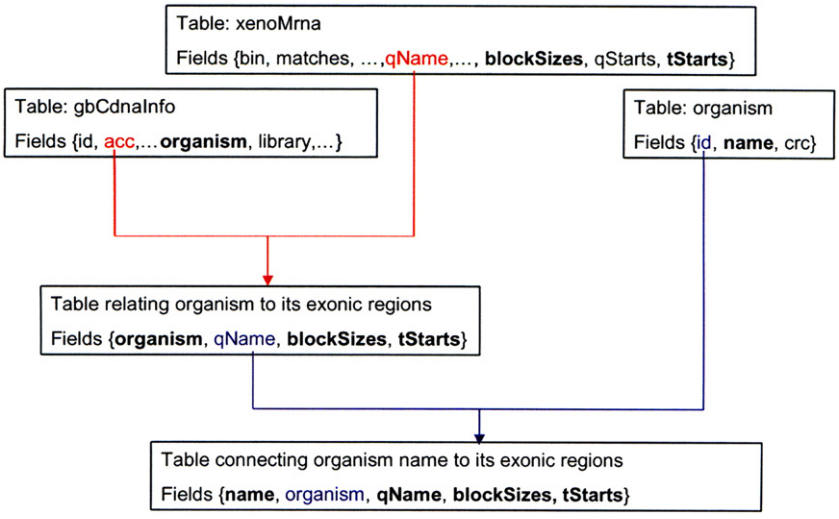

Fig. 3. Relating GenBank accession numbers of nonhuman cDNAs aligned with the human genome (xenoMrna) and species names from which the nonhuman cDNAs originated (organism). Red, blue, common fields used to join data together; bold, fields we are interested in.

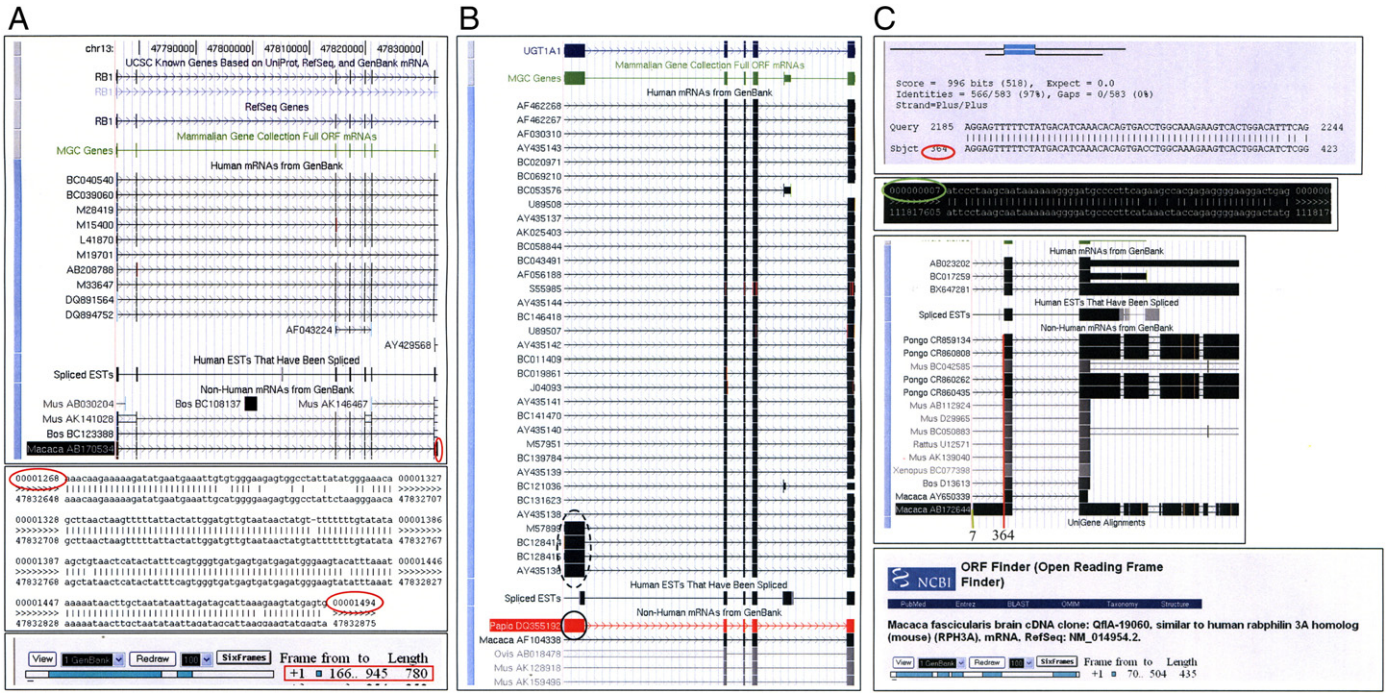

Fig. 4. Examples of unique terminal exons in nonhuman primate cDNA data: genomic localization and annotation. (A) Example of a unique primate-specific terminal exon not found in any human mRNAs from the orthologous locus. *Note.* The terminal exon starting and ending positions are at nucleotides (nt) 1268 and 1494, respectively (circled, top and middle panels). As the terminal exon lies after the largest positive-strand ORF (opening reading frame) of 166–945 (boxed, lower panel), we conclude that its uniqueness affects slowly the 3′UTR. (B) Example of implementing the majority-rule definition to characterize a terminal exon as primate-specific. We consider any nonhuman primate terminal exon as unique if it is absent from at least half of the human mRNAs. Here is a case of a unique primate-specific terminal exon (circled) that is present in the minority of human mRNAs (dotted circle). Categorization is similar to that of (A). (C) Example of a primate-specific terminal-exon extension. *Note.* From BL2SEQ of the macaque cDNA AB172644 and the orthologous human cDNA AB023202, we see that the stretch of identity between the macaque and the human cDNAs starts at nt 364 of the macaque cDNA (circled, top panel). Hence, the unique portion of the terminal exon is from nt 7 (circled, second panel from top) to nt 363 on the primate cDNA, a region that encompasses the 5′UTR (nt 7–69) and part of the ORF (nt 70–363; the complete ORF is nt 70–504) as revealed by NCBI ORF Finder. Thus, we concluded that the unique region, which is exonic in macaque but not in human in this locus, includes sequences from the 5′UTR and the CDS (but not the 3′UTR). The unique N-terminal fragment of the CDS translates to 98 unique amino acids. UCSC Genome Browser, NCBI BL2SEQ, and NCBI ORF Finder were used to construct the illustration.

human and nonhuman genomic localizations of gene boundaries, and vice versa). Human UTRs are generally <1 kb [49]. Therefore, differences between human genome coordinates of human and orthologous nonhuman cDNA ends that are less than 1 kb are assumed by us to refer to the same gene because their mapping discrepancy is less than the recognized size of untranslated regions. Unlike in orthologous transcript pairs with minimal differences at both ends, the genes identified by us are unlikely to match trivially the orthologous gene with perfect structural identity.

Identification of specific human–primate gene structure differences at orthologous loci

From human TU–primate xenoMrna orthologous pairs with overlapping boundaries, we eliminated all pairs that had boundary overlaps without exon overlaps. We then manually inspected all candidates for structural differences and categorized the unique nonhuman-primate-specific terminal exons as contributing 5′UTR, CDS, 3′UTR, or a combination of the three to the primate mRNAs they were a part of.

We favored manual inspection with the UCSC Browser due to the small number of loci of interest and due to the inability of computational pipelines to account correctly for all species differences observed. An unanticipated advantage of manual annotation is that we were able to identify several internal primate-specific exons and mRNAs that were intergenically spliced. We applied a stringent definition of primate-specific intergenic splicing. There had to be no human or nonprimate cDNA or EST support for each putative intergenic splice variant; the splice variant had to bridge two adjacent, known, nonoverlapping, annotated genes not connected by any human cDNAs or ESTs; and at least one existing splice site in each of the two genes, known from human cDNA/EST splicing analysis, had to be utilized by the primate intergenic splice variant. The intergenic splice sites were required to be GT–AG. Given these filtering criteria, the intergenic splices we uncovered most likely represent real transcriptional events. Bidirectional RACE/RT-PCR and resequencing in nonhuman primates would help confirm the reality and tissue specificity of these intergenic splicing events.

We had originally included human mRNAs whose accession numbers start with "CR". However, UCSC currently recognizes these accessions as unreliable. Therefore, we manually substituted human "CR" cDNA accessions with structurally most comparable "non-CR" alternatives where available.

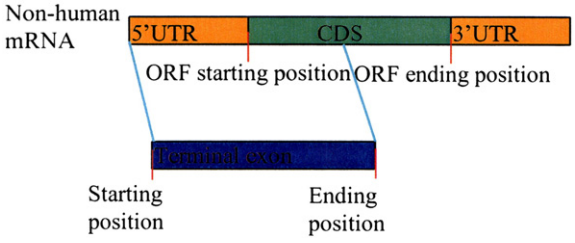

Fig. 5. Schematic illustration of UTR and CDS content determination for a primate-specific terminal exon. Lines show the relationship between the terminal exon and the UTR/CDS structure of the mRNA. The left line shows that base 1 of the terminal exon (first exon) is base 1 of the 5′UTR. The right line shows that the last base of the first exon is inside the CDS. Therefore, the terminal exon contains all of the 5′UTR and some of the CDS in this example.

The terminal exon of a nonhuman primate is defined as "unique" if it is absent from at least 50% of the human mRNAs (Fig. 4B). We observed several kinds of interspecies structural differences at gene ends.

Completely unique terminal exons

For this category of structural differences (Fig. 4A), we compared two sets of parameters: (1) the starting and ending positions of the terminal exon (obtained from UCSC Browser) and (2) the starting and ending position of the ORF (obtained from NCBI ORF Finder).

If the starting position of the terminal exon lay outside the range of the ORF and before the ORF starting position, but the ending position of the terminal exon lay within the ORF range, we concluded that the 5′UTR and part of the CDS belonging to the nonhuman mRNA were unique (Fig. 5).

If the start of the species-specific unique portion of the terminal exon in a nonhuman primate was inside the ORF, but the end was after the stop codon of the ORF, then we concluded that part of the CDS and 3′UTR belonging to the nonhuman mRNA were unique. If the unique species-specific terminal exon was entirely within the ORF, we concluded that only a part of the CDS of the nonhuman mRNA, and not any UTR sequence of that mRNA, was unique. If the unique species-specific terminal exon was entirely outside of the ORF and was either before the start or after the end of the ORF, we concluded that either the 5′UTR or the 3′UTR of the nonhuman mRNA, respectively, was unique.

Partially unique terminal exons

In this category, a nonhuman primate cDNA, when mapped to the human genome, shows that the primate ortholog of the human gene has a longer terminal exon relative to the human gene's structure. Therefore, only a contiguous portion of the terminal exon transcribed in the primate, and not the entire primate terminal exon, aligns to genomic sequence transcribed in human. We used NCBI BL2SEQ [50] in every case to determine the exact coordinates of the portion of the nonhuman primate cDNA that did not align with the orthologously encoded human cDNA (Fig. 4C).

Understanding the effect of structural differences on biological pathways, functions, and processes

To find out whether gene structure differences between humans and other primates affect specific biological pathways, functions, and processes, we uploaded the human accession numbers from the 183-gene total list and the 55-gene super-high-confidence list to www.pantherdb.org and three GO-based tools: David, FuncAssociate (beta version), and Genecodis. We compared the accession numbers of the human genes against both the "Human AB 1700 genes (REF)" and the "NCBI: H. sapiens genes (REF)" PantherDB lists.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.05.006.

References

[1] A. Varki, T.K. Altheide, Comparing the human and chimpanzee genomes: searching for needles in a haystack, Genome Res. 15 (2005) 1746–1758.
[2] The Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome, Nature 437 (2005) 69–87.
[3] Z. Zhang, A.W.C. Pang, M. Gerstein, Comparative analysis of genome tiling array data reveals many novel primate-specific functional RNAs in human, BMC Evol. Biol. 7 (Suppl. 1) (2007) S14.
[4] P. Stankiewicz, Serial segmental duplications during primate evolution result in complex human genome architecture, Genome Res. 14 (2004) 2209–2220.
[5] S.S.R. Datta. Gene hunters—sifting through evolution's shadows. Berkeley Science Review 2006 [cited]; available at www.sciencereview.berkeley.edu/articles/issue5/briefs_4.pdf.
[6] B.K. Dennehey, D.G. Gutches, E.H. McConkey, K.S. Krauter, Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution, Genomics 83 (2004) 493–501.
[7] T. Angata, E.H. Margulies, E.D. Green, A. Varki, Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms, Proc. Natl. Acad. Sci. USA 101 (2004) 13251–13256.
[8] Hayakawa, et al., A human specific gene in microglia, Science 309 (2005) 1693.
[9] M. Hurles, Gene duplication: the genomic trade in spare parts, PLoS Biol. 2 (2004) E206.
[10] Z. Cheng, et al., A genome-wide comparison of recent chimpanzee and human segmental duplications, Nature 437 (2005) 88–93.
[11] R. Sorek, G. Ast, D. Graur, Alu-containing exons are alternatively spliced, Genome Res. 12 (2002) 1060–1067.
[12] D.A. Ray, M.A. Batzer, Tracking Alu evolution in New World primates, BMC Evol. Biol. 5 (2005) 51.
[13] T. Hayaka, Y. Satta, P. Gagneux, A. Varki, N. Takahata, Alu-mediated inactivation of the human CMP-N-acetylneuraminic acid hydroxylase gene, Proc. Natl. Acad. Sci. USA 98 (1998) 11399–11404.
[14] R. Nielsen, et al., A scan for positively selected genes in the genomes of humans and chimpanzees, PLoS Biol. 3 (2005) E170.
[15] M.C. King, A.C. Wilson, Evolution at two levels in humans and chimpanzees, Science 188 (1975) 107–116.
[16] G. Glazko, et al., Eighty percent of proteins are different between humans and chimpanzees, Gene 346 (2004) 215–219.
[17] J.S. Mattick, et al., The functional genomics of noncoding RNA, Science 309 (2005).
[18] S. Katayama, et al., Antisense transcription in the mammalian transcriptome, Science 309 (2005) 1564.
[19] A.T. Willingham, et al., A strategy for probing the function of noncoding RNAs finds a repressor of NFAT, Science 309 (2005) 1570.
[20] K.C. Pang, M.C. Frith, J.S. Mattick, Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function, Trends Genet. 22 (2006).
[21] J.S. Mattick, Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms, Bioessays 25 (2003) 930–939.
[22] S.B. Carroll, Evolution at two levels: on genes and form, PLoS Biol. 3 (2005) e245.
[23] L.M. Steinmetz, L.W. Davis, Maximizing the potential of functional genomics, Nat. Rev. Genet. 5 (2004) 190–201.
[24] P. Khaitovich, et al., Regional patterns of gene expression in human and chimpanzee brains, Genome Res. 14 (2004) 1462–1473.
[25] A. Courseaux, J.L. Nahon, Birth of two chimeric genes in the Hominidae lineage, Science 291 (2001) 1293–1297.
[26] A. Ludwig, T.S. Rozhdestvensky, V.Y. Kuryshev, J. Schmitz, J. Brosius, An unusual primate locus that attracted two independent Alu insertions and facilitates their transcription, J. Mol. Biol. 350 (2005) 200–214.
[27] J.L. Nahon, Birth of 'human-specific' genes during primate evolution, Genetica 118 (2003) 193–208.
[28] P.H. Silverman, When is a gene a transcription unit? University of California at Irvine; published online at http://www.ethicscenter.uci.edu/pdf_documents/Paul%20Silverman%20Paper.pdf.
[29] J.C. Venter, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.
[30] M.D. Adams, et al., Complementary DNA sequencing: expressed sequence tags and human genome project, Science 252 (1991) 1651–1656.
[31] The FANTOM Consortium, The transcriptional landscape of the mammalian genome, Science 309 (2005) 1159.
[32] P.G. Engstrom, et al., Complex loci in human and mouse genomes, PLoS Genet. 2 (2006) e47.
[33] D. Karolchik, et al., The UCSC Genome Browser Database, Nucleic Acids Res. 31 (2003) 51–54.
[34] A. Varki, Nothing in glycobiology makes sense, except in the light of evolution, Cell 126 (2006) 841–845.
[35] X.S. Puente, et al., Comparative analysis of cancer genes in the human and chimpanzee genomes, BMC Genomics 7 (2006) 15.
[36] P.D. Thomas, et al., PANTHER: a library of protein families and subfamilies indexed by function, Genome Res. 13 (2003) 2129–2142.
[37] D.W. Huang, et al., DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acids Res. 35 (Web server issue) (2007) W169–W175.
[38] P. Carmona-Saez, M. Chagoyen, F. Tirado, J.M. Carazo, A. Pascual-Montano, GENECODIS: a Web-based tool for finding significant concurrent annotations in gene lists, Genome Biol. 8 (2007) R3.
[39] G.F. Berriz, O.D. King, B. Bryant, C. Sander, F.P. Roth, Characterizing gene sets with FuncAssociate, Bioinformatics 19 (2003) 2502–2504.
[40] B. Ma, R. Nussinov, Trp/Met/Phe hot spots in protein–protein interactions: potential targets in drug design, Curr. Top. Med. Chem. 7 (2007) 999–1005.
[41] K. Maeda, T. Horikoshi, E. Nakashima, Y. Miyamoto, MATN and LAPTM are parts of larger transcription units produced by intergenic splicing: intergenic splicing may be a common phenomenon, DNA Res. 12 (2005) 365–372.
[42] P.G. Zaphiropoulos, RNA molecules containing exons originating from different

members of the cytochrome P450 2C gene subfamily (CYP2C) in human epidermis and liver, Nucleic Acids Res. 27 (1999) 2585–2590.
[43] I.E. Gallouzi, et al., HuR binding to cytoplasmic mRNA is perturbed by heat shock, Proc. Natl. Acad. Sci. USA 97 (2000) 3073–3078.
[44] P. Lasko, Gene regulation at the RNA layer: RNA binding proteins in intercellular signaling networks, Sci. STKE 2003 (2003) re6.
[45] L. He, G.J. Hannon, MicroRNAs: small RNAs with a big role in gene regulation, Nat. Rev. Genet. 5 (2004) 522–531.
[46] A. Clop, et al., A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep, Nat. Genet. 38 (2006) 813–818.
[47] M. Hiller, R. Pudimat, A. Busch, R. Backofen, Using RNA secondary structures to guide sequence motif finding towards single-stranded regions, Nucleic Acids Res. 34 (2006) e117.
[48] W. Liu, S.Y. Mao, W.Y. Zhu, Impact of tiny miRNAs on cancers, World J. Gastroenterol. 13 (2007) 497–502.
[49] F. Lopez, S. Granjeaud, T. Ara, B. Ghattas, D. Gautheret, The disparate nature of "intergenic" polyadenylation sites, RNA 12 (2006) 1794–1801.
[50] T.A. Tatusova, T.L. Madden, BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences, FEMS Microbiol. Lett. 174 (1999) 247–250.
[51] N. Osada, et al., Substitution rate and structural divergence of 5′UTR evolution: comparative analysis between human and cynomolgus monkey cDNAs, Mol. Biol. Evol. 22 (2005) 1976–1982.